Inferred Developmental Origins of Brain Tumors from Single Cell RNA-Sequencing Data

Su Wang [1,+], Rachel Naomi Curry, B.A.[2,3,+], Anders W Erickson[4,5,6], Claudia Kleinman[7,8], Michael D. Taylor, M.D.[9,10], Ganesh Rao, M.D.[1,11], Benjamin Deneen, Ph.D. [1,3,11], Arif O. Harmanci, Ph.D. [12,‡], Akdes Serin Harmanci, Ph.D. [1,11,‡]

[1]Department of Neurosurgery, Baylor College of Medicine, Houston, TX, USA

[2]Graduate School of Biomedical Sciences, Baylor College of Medicine, Houston, TX, USA

[3]Center for Cell and Gene Therapy, Baylor College of Medicine, Houston, TX, USA

[4]The Arthur and Sonia Labatt Brain Tumor Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada

[5]Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada

[6]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada

[7]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada

[8]Department of Human Genetics, McGill University, Montreal, Quebec, Canada

[9]Texas Children's Cancer Center, Hematology-Oncology Section, Houston, TX, USA

[10]Department of Pediatrics - Hematology/Oncology and Neurosurgery, Baylor College of Medicine, Houston, TX, USA

[11]Center for Cancer Neuroscience, Baylor College of Medicine, Houston, TX, USA

[12]McWilliams School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA

[+] These authors contributed equally to this work

[‡] Corresponding Author

Abstract

The reactivation of neurodevelopmental programs in cancer highlights parallel biological processes that occur in both normal development and brain tumors. Achieving a deeper understanding of how dysregulated developmental factors play a role in the progression of brain tumors is therefore crucial for identifying potential targets for therapeutic interventions. Single-cell RNA sequencing (scRNA-Seq) provides an opportunity to understand how developmental programs are dysregulated and reinitiated in brain tumors at single-cell resolution. Here, we introduce COORS (Cell Of ORigin like CellS), as a computational tool trained on developmental human brain single-cell datasets that enables annotation of “developmental-like” cell states in brain tumor cells. Applying COORS to various brain cancer datasets, including medulloblastoma (MB), glioma, and diffuse midline glioma (DMG), we identified developmental-like cells that represent putative cells of origin in these tumors. Our work adds to our cumulative understanding of brain tumor heterogeneity and helps pave the way for tailored treatment strategies.

Introduction

One of the greatest challenges to finding a cure for brain cancers is the robust inter- and intra-tumoral heterogeneity that characterizes these tumors[1–4]. This heterogeneity contributes to disease progression and is a key reason therapeutic approaches fail to prevent disease recurrence. Although the genetic evolution of cancer cells is a critical determinant, tumor heterogeneity is also influenced by non-genetic factors including varying developmental cellular programs, which include stem, progenitor, and senescent cell states[5,6]. Prior studies have demonstrated that aberrant expression of neurodevelopmental programs is pervasive in brain tumors and is largely driven by the reactivation of developmental transcriptional states that are acquired by genomic and epigenomic changes. Given the complexity of cell types and an array

of developmental states, isolating a single cell type of origin poses a difficult task, however, a more thorough examination of brain tumor transcriptomics alongside transcriptional signatures of neurodevelopmental cell types may shed light on the origins of brain cancer. To gain a deeper understanding of which developmental cell types brain tumors most closely resemble, we hypothesized that tumor cell lineages can recapitulate cell lineages encountered in the developing brain. While tumors exhibit a multitude of dysregulated pathways, existing evidence, particularly in pediatric tumors, supports this hypothesis[6–10]. We, therefore, focused on employing developmental expression modeling trained on human brain atlases that span various developmental time points. This modeling approach allows us to characterize tumor cells by overlaying their gene expression patterns onto those of early neurodevelopmental stages. By identifying and studying these myriad cellular states from development, our goal is to uncover insights into the origins and behavior of brain tumors, ultimately paving the way for more effective treatment strategies and improved patient outcomes.

Single-cell RNA sequencing (scRNA-Seq) provides an opportunity to dissect the complex cellular states during development and in health and disease[11]. However, it is computationally challenging to decipher the spectrum of heterogeneous developmental cell states in tumor cells using scRNA-Seq. Accurate identification of developmental-like cell states necessitates a comprehensive understanding of the interactions among all genes, which, in turn, requires a substantial amount of gene expression data. In this study, we developed COORS, a computational tool to annotate each developmental cell state in tumor cells at single cell resolution. COORS uses multilayer perceptron model for cell of origin classification and cell age regression using developing brain scRNA-Seq datasets from previously published scRNA-seq datasets, comprising approximately 1M million cells from developing human and mouse brains[6,12–16]. We used COORS to predict developmental-like analogs in pediatric and adult

tumors using public and in-house MB[17], DMG[14] and glioma scRNA-Seq data, which revealed unique developmental cell types as putative cells of origin in each brain cancer subtype.

Results

The overall workflow of COORS algorithm

The basic motivation of COORS algorithm is to utilize existing primitive cell types identified in developing normal brain tissues to annotate the potential origins of cells in tumor samples. In this regard, COORS aims to capitalize on the large amount of healthy developmental transcriptional profiles to identify similar programs in tumor tissues. This is in essence different from existing cell type annotation methods with two aspects: 1) Existing methods for cell type annotations in healthy tissues rely on exact matching of the cell types. 2) Existing methods for cell type annotations in tumor samples focus mainly on detecting tumor cells and do not focus on annotating the origin states for cells.

For training COORS models, we train a multilayer perceptron model for cell of origin classification and cell age regression using developing brain scRNA-Seq datasets[6,12–16] (**Table S1**). Assuming we have reference data with two origin-like cell types A and B, we train a neural network-based cell of origin classifier using this reference data, saving the model in our repository **(Figure 1A-B)**. Concurrently, we train two neural network-based cell age regressors, one for cell origin A and another for cell origin B, also saving these trained models in the repository. In the assignment step, each tumor needs to be matched to correct set of origin-like cell type assignment models. This is justified since each cancer originates at different regions of brain (neocortex vs cerebellum vs pons) which can be used to select the COORS models. After selecting the relevant origin-like cell models, we map tumor cells to developing healthy brain cells by using the pre-trained models. We predict the cell of origin for the testing dataset using the pre-trained cell of origin classifier. For each developmental-like cell type, we further predict

cell age using the corresponding pre-trained cell age regressors. Finally, we conduct SHapley Additive exPlanations (SHAP) analysis[18] to extract essential features from our machine-learning neural network models, identifying tumor-specific developmental-like gene markers for each cell type and age within our training dataset **(Figure 1A-B).**

Validation of COORS algorithm on scRNA-Seq Medulloblastoma data

Medulloblastoma (MB) is a pediatric brain tumor that is closely associated with early hindbrain development and can be classified into four main molecular subgroups[6,7,10,19–22]. The WNT-activated subgroup is defined by mutations in the WNT signaling pathway and generally displays a favorable prognosis. In contrast, the Sonic Hedgehog (SHH)-activated subgroup results from mutations in the SHH pathway and may have varying clinical outcomes. Group 3 (GP3) MBs have a distinct gene expression profile and are typically associated with a poorer prognosis. Finally, Group 4 (GP4) tumors, marked by a specific gene expression pattern, tend to have intermediate clinical outcomes. Understanding the origin of these four tumor subtypes might lead to the development of improved treatment strategies.

We applied the COORS algorithm to previously published MB scRNA-Seq data [17], where subgroup annotations are available for each sample, containing 29 samples and approximately ~40K cells in total (**Figure 2A**). Using COORS algorithm, we have used the pre-trained cell type and cell age models, derived from scRNA-Seq data of developing cerebellum, to map tumor MB cells (**Figure 2B**)[13]. We have not focused on WNT subgroup tumor cells because the WNT subgroup is known to originate from the lower rhombic lip (LRL) adjacent to the brainstem, rather than from the upper rhombic lip (URL) in the cerebellum[23]. COORS maps SHH subgroup tumor cells to granule cell precursor (GCP), GP4 subgroup tumor cells to unipolar brush cell (UBC-CN), Granule Neurons (GN) predominantly to SHH and secondarily to GP4 and GP3

subgroup tumor cells, and Rhombic Lip (RL) subgroup tumor cells to GP4 and GP3 (**Figure 2B-F**). These findings align with recent publications on MB subgroup cell origins[8,9,24,25]. Earlier research identified GN as progenitor cells for SHH-induced MB, while another study proposed that the rhombic lip subventricular zone (RLSVZ) progenitor cells are the source of GP3 and GP4 MB[8,9,24,25]. Additionally, one study suggested that GP4 originates from the UBC lineage[8]. In addition, we conducted SHAP analysis[18] to extract critical features from our machine-learning neural network models, which returned new and known marker genes associated with upper rhombic lip-derived cell types and the MB subgroups to which they correspond. For example, markers associated with the external granule layer or GCP identity such as *NDST3*[26], *CBFA2T2*[24], and *UNC13C*[27]; GC identity and maturation such as *RBFOX3*[28,29], *GRIK2*[24], *ROBO1* whose paralogs are essential in GC migration[30], with *MSI2* likely as it marks GCPs in the external granule layer but not postmitotic granule cells in the internal granule layer[31]; UBC identity such as *LMX1A*[32], *CACNA2D1*[8,33], *RELN*[34], known GP4 oncogene *ERBB4*[24,35] and *JMJD1C/KDM3C*, a putative H3K9 demethylase whose paralogs are recurrently somatically altered in GP4 MB[20,24,36]; and RL identity such as *OTX2*[13], *HIST1H4C, -3B, and -1C*[37], and *SLIT2*[30] (**Figure 2G-J**).

Next, we predicted the cell age of each identified developmental cell type within MB tumor cells using our pretrained cell age regressor models **(Figure 2K-O).** Interestingly, GP4 subgroup cells mapping to UBC mostly correspond to the later weeks in development by 17 post-conceptional weeks (PCW), while GP3 and GP4 subgroup cells mapping to RL mostly correspond to the earlier weeks by 11 PCW in development (**Figure 2K)**. The DEGs between tumor cells and their respective cells of origin, as well as the DEGs between the cell of origin and the following developmental stage cell type are listed in **Table S2 (Figure S1-8)**.

Application of COORS algorithm on pediatric diffuse midline glioma (DMG)

Next, we applied the COORS algorithm on previously published H3.1/H3.2 and H3.3 histone 3 K27M-mutant DMG scRNA-Seq data containing 13 samples and approximately 47K cells in total (**Figure 3**)[6,14]. Using COORS algorithm, similar to our previous application in MB data, we have used the pre-trained cell type and cell age models, derived from scRNA-Seq data of developing mouse pons brain[6], to map pediatric glioma cell origins (**Figure 3A-C**). Consistent with previous studies we have found that H3.3K27M gliomas mapped to oligodendrocyte-like and neuron-like cells, as previously reported[3,6,14,38]. More specifically, H3.1/2K27M tumors mapped to ependymal-like cells whereas H3.3 mapped to neuronal intermediate progenitor cells (IPCs).

Additionally, SHAP analysis identified that *FOXJ*, a well-known ependymal transcription factor[39], contributed to the mapping of H3.1/2 tumor cells to ependymal origins, while NFIB, a recognized transcription factor for neuronal progenitors[40], contributed to the assignment of H3.3 tumor cells to neuronal IPCs. Next, we predicted the cell age of each identified developmental cell type within DMG tumor cells using our pretrained cell age regressor models **(Figure S9).** H3.3 tumor cells mapping to neuronal IPCs mostly correspond to the earlier weeks in development, whereas H3.1/2 tumor cells mapping to ependymal-like cells mostly correspond to the later weeks in development (**Figure S9**). The DEGs between tumor cells and their respective cells of origin, as well as the DEGs between the cell of origin and the subsequent developmental cell type, are listed in **Table S3 (Figure S10-27)**.

Application of COORS algorithm on inhouse scRNA-Seq glioma data

Next, we applied the COORS algorithm to our inhouse glioma scRNA-Seq data, containing 21 samples and approximately ~234K cells in total (**Figure 4A**)[41]. Using COORS algorithm, we have used the pre-trained cell type and cell age models, derived from Jessa et al. developing mouse forebrain[6] and three human developing brain scRNA-Seq datasets from Zeng et al[15],

Polioudakis et al[128] and Bhaduri et al[16] to map adult glioma cells (**Figure 4B-D**). Pretrained models from both mouse forebrain developmental dataset[12] and the dataset by Bhaduri et al.[41,42] maps IDH^{WT} subgroup tumor cells to the radial glia (RG) cells. RG cells are neural stem cells found in the developing human brain, particularly during the embryonic stages of brain development[42,43] and alterations in their regulatory pathways or genetic mutations can lead to their transformation into glioma stem-like cells (GSCs), which are thought to drive tumor initiation and progression in glioblastoma[44–46]. In another dataset on the developing human brain by Zeng et al.[15], IDH^{WT} tumor cells are again mapped to Neural Stem Cells (NSC-cluster 12). Poliodakis et al.[41,42] dataset more specifically maps IDH^{WT} tumor cells to ventricular zone radial glia (vRG) which are known to reside in the ventricular zone (VZ) of the developing brain[42,43]. On the other hand, IDH^{Mut} cells consistently maps to oligodendrocytes, OPC and also to neuronal subtypes using multiple pretrained models from Bhaduri et al.[16], Poliodakis et al.[12], Jessa et al.[6] datasets. In another dataset on the developing human brain by Zeng et al.[46,47], IDH^{Mut} tumor cells mostly mapped to specific populations of GABA cells in the developing brain (GABA-cluster 9) derived from the ectoderm layer. Previous studies also suggest that GABAergic neurons and OPCs are derived from common neurodevelopmental origins; predominantly, they both originate from Nkx2.1-expressing precursors located in the medial, lateral, and caudal ganglionic eminences[47,48]. Moreover, GABAergic neurons and OPCs converge at a shared transcriptional state with expression of *OLIG2*[49], *GABARs*[50], and *PDGFRA*[51]. In our recent study, we demonstrated that a subset of IDH^{Mut} glioma cells fire single, short action potentials (APs) and are defined by mixed characteristics of GABAergic neurons and OPC[52].

In addition, we conducted SHAP analysis to extract critical features from our machine-learning neural network models, enabling the identification of developmental-like gene markers specific to glioma for each mapped cell type. SHAP analysis identified that markers commonly

associated with oligodendrocyte precursor cells (OPCs) such as *OLIG1*[53], *SOX10*[54], *PDGFRA*[51] *NKX2-2*[55] and *OLIG2*[56,57] predominantly contributed to the mapping of IDH mutant tumor cells to OPCs. GABAA receptor subunit *GABRA2*[58], known to be also expressed in oligodendrocytes contributed to the mapping of neuronal subtypes in IDH^{Mut} using Jessa et al.[6] pretrained model. In the dataset from Zeng et al.[15], GABA-cluster 9 exhibits a high expression of genes such as *STMN2*, *ELAVL4*, *ELAVL3*, and *DCX*, all of which are known to have crucial role in neuronal development and differentiation, predominantly contributed to the mapping of IDH^{Mut} tumor cells to neuronal subtypes. SHAP analysis identified that markers commonly associated with neural stem cells and radial glia cells *HES5*[59]*and HES1*[59], contributed to the mapping of IDH^{WT} tumor cells to radial glia cells. NSC-cluster 12 exhibits high expression of genes such as *SOX2*, and *TTYH1*, all of which are known markers of neural stem cells[60,61], contributing to the mapping of IDH^{WT} tumor cells to NSCs.

To estimate the cell age of each identified developmental cell type within glioma tumor cells, we applied our pre-trained cell age regressor models (**Figure 4A-C**). Age mapping was performed exclusively on the datasets from Bhaduri et al.[16] , Zeng et al.[15] , and Jessa et al.[6] due to their wide range of developmental age data. Notably, both Bhaduri et al.[16] and Jessa et al.[6] datasets' models consistently revealed that IDH^{Mut} cells corresponding to OPCs exhibit a bimodal age distribution, indicating stages early and late in development. In contrast, IDH^{WT} cells aligning with RG mostly represent earlier developmental weeks. The DEGs between tumor cells and their respective cells of origin, as well as the DEGs between the cell of origin and the following developmental cell type, are listed in **Table S4-5 (Figure S28-51)**.

Discussion

Here we presented a hierarchical machine learning-based approach, named COORS, for the identification and characterization of tumor cells that exhibit gene expression patterns reminiscent of early developmental stages in the brain. COORS achieves this by employing NNMs trained on diverse scRNA-Seq datasets from developing human brain tissues. We applied our method to predict developmental-like cells in various brain cancer datasets, including MB, DMG, and glioma, with validation against well-characterized MB data. COORS identified vRG developmental cells within IDH^{WT} glioma cells whereas OPC and neuronal-like cells in IDH^{Mut}. Interestingly, IDH^{Mut} subgroup cells that map to OPC show bimodal distributions, that are both early and late weeks in development, while IDH^{WT} subgroup cells mapping to RG mostly correspond to the earlier weeks in development. Furthermore, COORS offers a valuable resource by providing information on the DEGs between tumor cells and their respective cells of origin, as well as between the cell of origin and the subsequent developmental cell type. These DEGs hold promise as potential therapeutic targets, offering new avenues for the development of targeted therapies for brain tumors. In conclusion, the development and application of COORS represent a significant advancement in our ability to accurately annotate developmental-like cell states in brain cancer datasets and potentially extend this approach to other cancer types.

In the past, efforts to induce the differentiation of cancer cells into more mature, less aggressive cell types, without damaging normal cells were met with limited success in solid tumors, likely due to insufficient understanding of the precise progenitor cells involved[62,63]. However, advancements in our comprehension of specific time points and cell types have paved the way for a more nuanced approach. By examining the subsequent steps in the lineage, such as the differentiation of OPCs into mature oligodendrocytes, we can identify key genes involved in this process. For instance, oligodendrocytes are characterized by decreased proliferative capacity compared to OPCs, suggesting a regulatory role for certain genes in cell fate determination. Targeting genes like *OLIG2*, which maintains OPC identity, while activating those

involved in mature oligodendrocyte function, such as myelin basic protein (*MBP*), holds promise for directing OPCs toward differentiation into oligodendrocytes. This approach presents a potential avenue for differentiation therapy, wherein manipulating gene expression could drive tumor cells toward a more benign phenotype, offering a novel strategy for cancer treatment.

Overall, our approach relies on a separate model for each origin-like cell type. While one could use a single model for assignment of all cell types jointly (e.g. multitask learning) to make use of all data at once, our experiments did not show significant benefit of one model compared to building simpler single models for each cell type. We hypothesize that this may be inherently a result of data size requirements and complexities of jointly learning hundreds of origin-like cell types from unbalanced datasets. In addition, our approach provides more flexibility for selecting biologically meaningful models of origin-like cell types in different tumors.

Methods

Data preprocessing

We conducted the standard pipeline of single-cell RNA sequencing data preprocessing for both reference and testing datasets using Scanpy 1.7.2 in Python 3.6.8. For cell of origin classification, the preprocessing of reference and testing data were performed starting from the whole datasets. On the other hand, for cell age regression, we first grouped the reference and testing datasets by cells of origin and then preprocessed each group separately. Each cell was normalized to have the same total read count and the matrices were transformed into natural logarithm domain. We annotated the top 2,000 highly variable genes in reference dataset, scaled both datasets to unit variance and zero mean, and truncated to 10. We kept all the other parameters in default values.

A common gene set is needed between the reference and testing datasets as they will be fed into one same model. We took the union of reference marker genes and highly variable genes to intersect with the testing genes as the common gene set and trimmed both datasets. The volumes of reference cell types vary. We first excluded cell types with number of cells fewer than 20. To balance the number of cells among the rest cell types, we set as baseline the cell type that is the 25% quantile in terms of numbers of cells and randomly subsampled the others to this baseline. Those cell types with number of cells fewer than the baseline were not subsampled.

We randomly split each reference cell type into two subsets, one with 80% cells for model training and the other with 20% cells for model validation. The training subsets of all the cell types were concatenated as training data, the validation subsets as validation data. We conducted one-hot encoding of cell types in both training and validation data. We scaled training, validation, and testing data into the range from 0 to 1 using Scikit-learn 0.24.2.

MLP-based prediction model

We developed multilayer perceptron networks for cell of origin classification and cell age regression.

Cell of origin classifier. The cell of origin classifier has one input layer, variable numbers of hidden layers, and one output layer. The input layer has the same number of nodes as the input genes. The number of hidden layers varies from one to four, and the number of nodes in one hidden layer is set to be 256, 128, 64, or 32, which is determined after hyperparameter optimization. Following the dense connection within each hidden layer, there are batch normalization, activation, and dropout functions. We use the popular Rectified Linear Unit (ReLU) for hidden layer activation and set dropout rate to be 0.1 or 0.2. The output layer uses Softmax activation function so that each node outputs a non-negative value smaller than 1 and

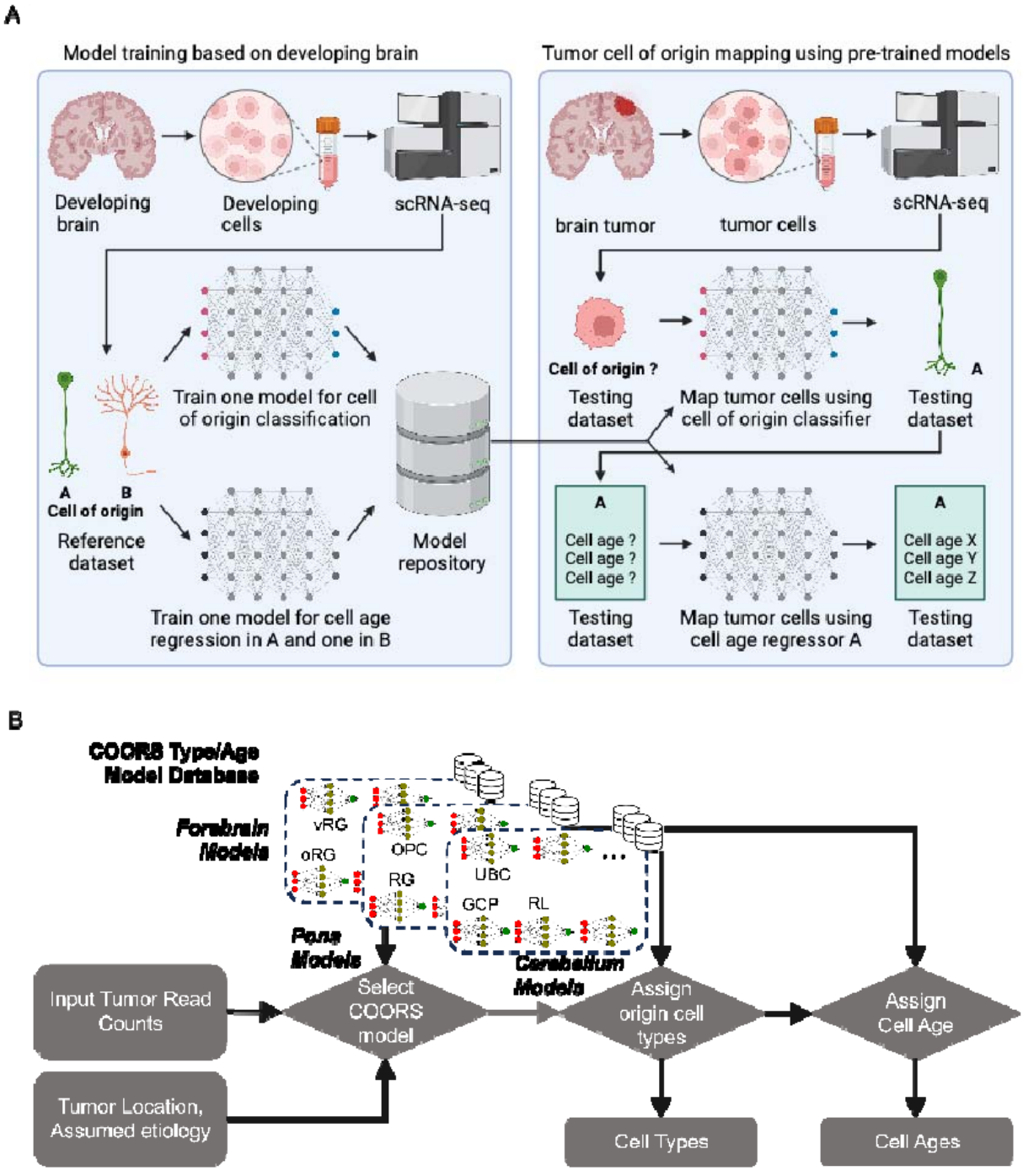
all the values sum up to 1. Therefore, each output corresponds to the probability of one cell type. We compile the model using categorical crossentropy as loss function, Adam as optimizer, and accuracy as metrics.

Cell age regressor. Similar as cell of origin classifier, cell age regressor consists of one input layer, a group of hidden layers, and one output layer. While the input layer and hidden layers are structurally the same as cell of origin classifier, the output layer of cell age regressor has only one node with Sigmoid activation function that corresponds to the predicted cell age. The model is compiled using mean squared error as loss function, Adam as an optimizer, and loss as metrics. Since more than one cell age regressors exist corresponding to each cell of origin classifier, these regressors can have specific hyperparameters of hidden layers that are not necessarily the same.

Model training prerequisites. We implemented cell of origin classifier and cell age regressor using Keras 2.6.0 with Tensorflow 2.6.2 as backend in Python 3.6.8. Prerequisite packages for data preprocessing and model training include Numpy 1.19.5, Pandas 1.1.5, Scanpy 1.7.2, Anndata 0.7.8, Scipy 1.5.4, and Scikit-learn 0.24.2.

Hyperparameter optimization. We systematically optimized hyperparameters of cell of origin classifier and cell age regressor using grid search cross validation implemented by Scikit-learn 0.24.2, focusing on tuning the number of hidden layers and nodes, dropout rate of hidden layers, and learning rate of the optimizer. For each model, we varied the number of hidden layers from one to four and the number of nodes in each layer that could be 256, 128, 64, or 32. We followed convention to use Rectified Linear Unit (ReLU) as activation function in hidden layers. Training epochs were fixed to be 100 and batch size 32 as they did not show significant affections in our case. Along with the iteration of every possible hidden layer structure, we

334 explored dropout rate of 0.1 or 0.2 and learning rate of 0.1, 0.01, 0.001, or three decaying
335 learning rates that were scheduled to exponentially reduce during model training based on initial
336 rate of 0.1 or 0.01, final rate of 0.01 or 0.001, training epochs, and batch size.
337
338
339 **Source Code Availability**
340 Source code of COORS is publicly available at https://github.com/Su-Wang-UTH/COORS



341

342 ***Figure 1. Overview of COORS algorithm.*** (A) In the first step, neural network models are
343 trained for cell of origin classification and cell age regression using developing brain scRNA-Seq
344 datasets, and the models saved in the repository. In the second step, these pre-trained models

are used to map scRNA-Seq tumor cells to developing brain cells, predicting cell origin and age while conducting SHAP analysis to identify tumor-specific gene markers. (B) Tumors are matched with specific origin-like cell type assignment models based on their region of origin within the brain (e.g., neocortex vs cerebellum vs pons), enhancing the precision of the COORS application. Post-model selection, the mapping of tumor cells to developing healthy brain cells is performed through the application of these pre-trained models, as depicted in the schematic.

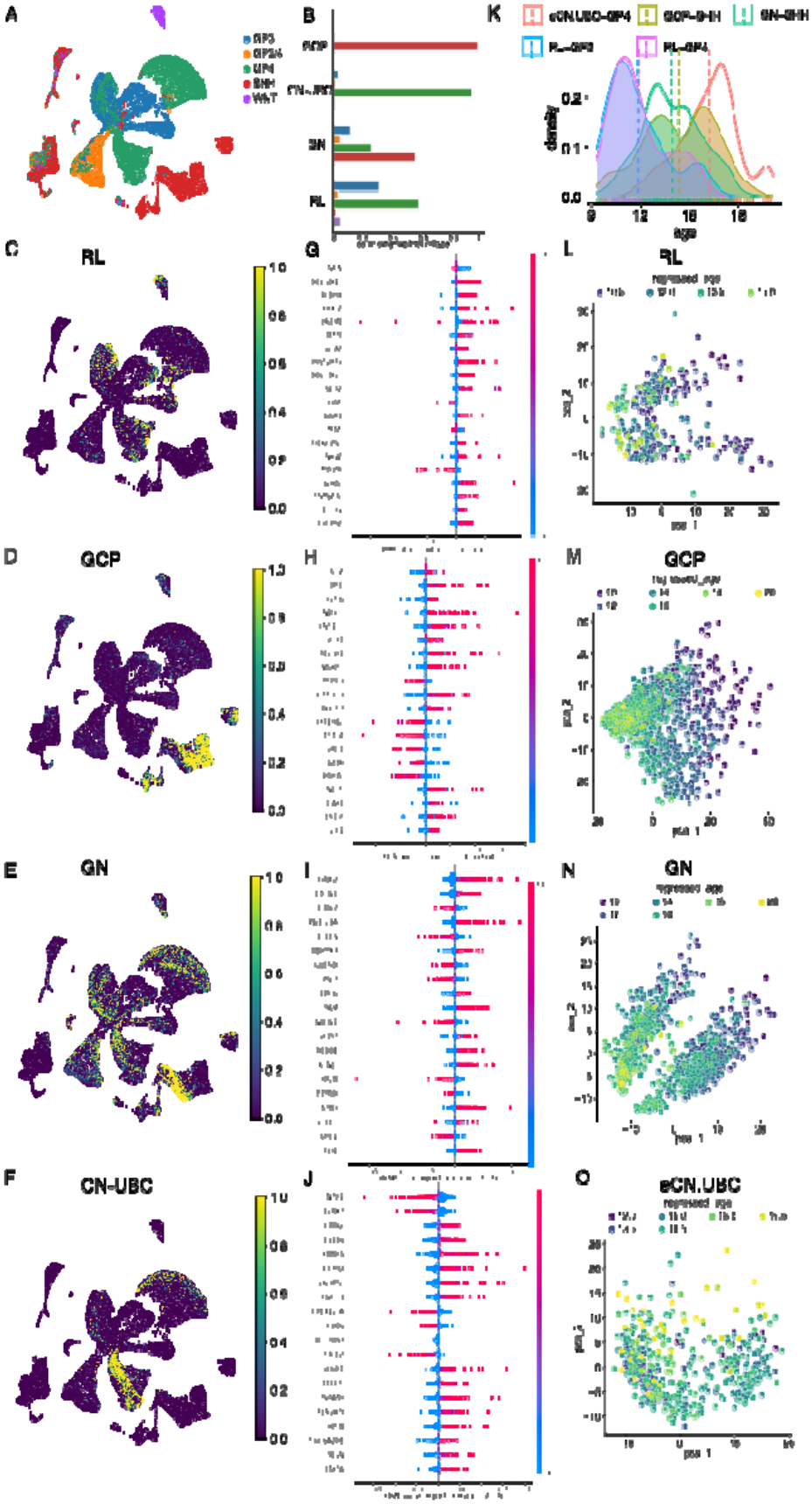


Figure 2. Characterization of developmental-like cell states in MB scRNA-Seq data. (A) Tumor subgroups are shown for MB scRNA-seq dataset. (B) Barplots showing the distribution of MB cells within each tumor subgroup mapped to individual developing cell types. (C-F) Developmental cell type probability scores are shown for RL, GCP, GN and UBC cell types. (G-J) The figure displays the results of SHAP analysis, showing the top impactful genes from each cell type, RL, GCP, GN and UBC respectively, in our training dataset. (K) Distribution of age mapping within each tumor subgroup and their respective mapped cell of origin pairs. (L-O) PCA plots showing the mapping of developmental ages for tumor cells mapped to various developmental origins, GCP, CN-UBC, GN, and RL respectively.

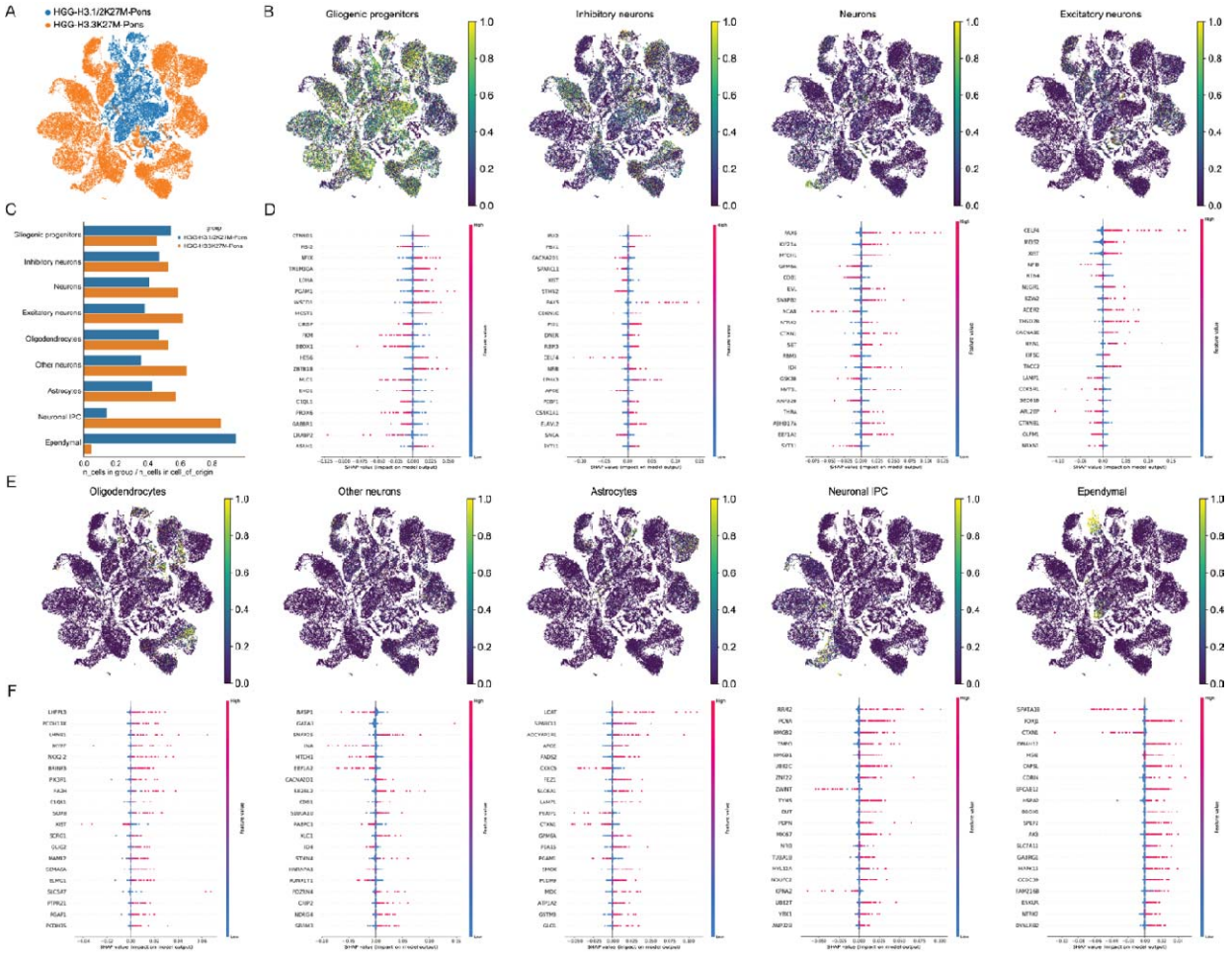
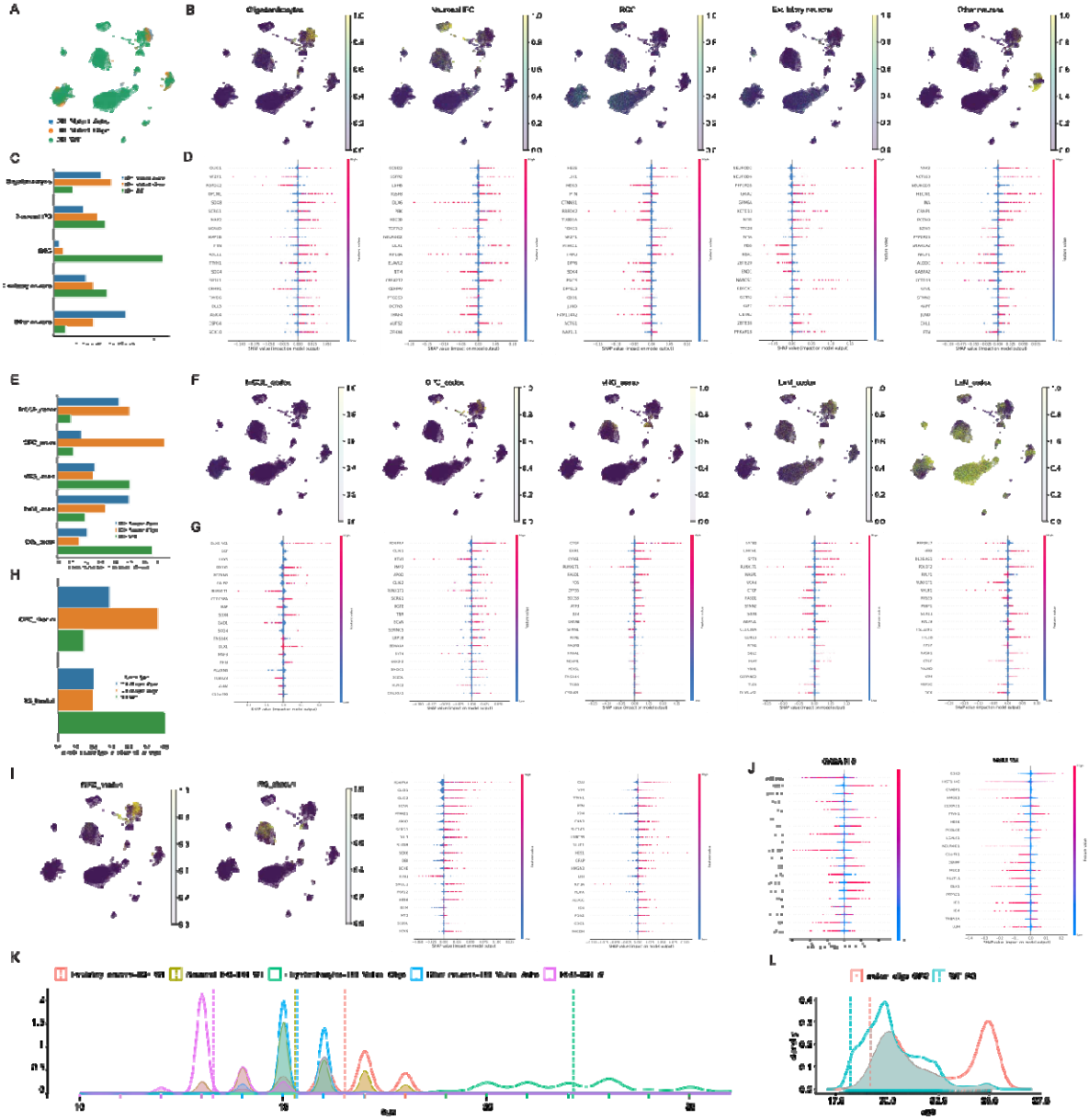


Figure 3. Characterization of developmental-like cell states in DMG scRNA-Seq data. **(A)** Tumor subgroups are shown for DMG scRNA-seq dataset. **(B)** Developmental cell type probability scores are shown for Gliogenic Progenitor, Inhibitory Neurons, Neurons, and Excitatory Neurons cell types. **(C)** Barplots showing the distribution of DMG cells within each tumor subgroup mapped to individual developing cell types. **(D)** The figure displays the results of SHAP analysis, showing the top impactful genes from each cell type Gliogenic Progenitor, Inhibitory Neurons, Neurons, and Excitatory Neurons respectively, in our training dataset. **(E)** Developmental cell type probability scores are shown for Oligodendrocytes, Other neurons, Astrocytes and Neuronal IPC, and Ependymal cell types **(F)** The figure displays the results of SHAP analysis, showing the top impactful genes from each cell type Oligodendrocytes, Other neurons, Astrocytes and Neuronal IPC and Ependymal respectively, in our training dataset.

386



387

388 ***Figure 4. Characterization of developmental-like cell states in glioma scRNA-Seq data.***

389 **(A)** Tumor subgroups are shown for inhouse glioma scRNA-seq dataset. **(B)** Developmental

390 cell type probability scores are shown for Oligodendrocytes, Neuronal IPC, RGC, Excitatory

391 Neurons and Other neurons predicted from Jessa et al pretrained models[14]. **(C)** Barplots

392 showing the distribution of glioma cells within each tumor subgroup mapped to individual

393 developing cell types in Jessa et al dataset[14]. **(D)** The figure displays the results of SHAP

analysis, showing the top impactful genes from each cell type, Oligodendrocytes, Neuronal IPC, RGC, Excitatory Neurons and Other neurons respectively, using Jessa et al pretrained models[14]. **(E)** Barplots showing the distribution of glioma cells within each tumor subgroup mapped to individual developing cell types predicted from Poliudokis et al. pretrained models[12]. **(F)** Developmental cell type probability scores are shown for InCGE (caudal ganglionic eminence derived interneurons), OPC, vRG, ExM (migrating excitatory neuron) and ExN (newborn excitatory neuron) cell types predicted from Poliudokis et al. pretrained models[12]. **(G)** The figure displays the results of SHAP analysis, showing the top impactful genes from each cell type in InCGE, OPC, vRG, ExM and ExN celltypes using Poliudokis et al. pretrained models[12]. **(H)** Barplots showing the distribution of glioma cells within each tumor subgroup mapped to individual developing cell types in Bhaduri et al. dataset[16]. **(I)** Developmental cell type probability scores are shown for OPC, RG cell types in Bhaduri et al. dataset[16] and SHAP analysis, showing the top impactful genes from each cell type, OPC, RG respectively, using Bhaduri et al. pretrained models[16]. **(J)** The figure displays the results of SHAP analysis, showing the top impactful genes from each cell type, GABA N 9 and NSC 12 respectively, using Zeng et al. pretrained models[15]. **(K)** Distribution of age mapping within each tumor subgroup and their respective mapped cell of origin pairs from Jessa et al pretrained models[14]. **(L)** Distribution of age mapping within each tumor subgroup and their respective mapped cell of origin pairs from Bhaduri et al. pretrained models[16].

References

1. Friedmann-Morvinski, D. Glioblastoma heterogeneity and cancer cell plasticity. *Crit. Rev. Oncog.* **19**, 327–336 (2014).
2. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
3. Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331–335 (2018).
4. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, (2017).
5. Curry, R. N. & Glasgow, S. M. The role of neurodevelopmental pathways in brain tumors. *Front. Cell Dev. Biol.* **9**, 659055 (2021).
6. Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nat. Genet.* **51**, 1702–1713 (2019).
7. Phoenix, T. N. The origins of medulloblastoma tumours in humans. *Nature* vol. 609 901–903 (2022).
8. Smith, K. S. *et al.* Unified rhombic lip origins of group 3 and group 4 medulloblastoma. *Nature* **609**, 1012–1020 (2022).
9. Selvadurai, H. J. *et al.* Medulloblastoma arises from the persistence of a rare and transient Sox2+ granule neuron precursor. *Cell Rep.* **31**, 107511 (2020).
10. Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* **572**, 67–73 (2019).
11. Jovic, D. *et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* **12**, e694 (2022).
12. Polioudakis, D. *et al.* A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785-801.e8 (2019).

13. Aldinger, K. A. *et al.* Spatial and cell type transcriptional landscape of human cerebellar development. *Nat. Neurosci.* **24**, 1163–1175 (2021).
14. Jessa, S. *et al.* K27M in canonical and noncanonical H3 variants occurs in distinct oligodendroglial cell lineages in brain midline gliomas. *Nat. Genet.* **54**, 1865–1880 (2022).
15. Zeng, B. *et al.* The single-cell and spatial transcriptional landscape of human gastrulation and early brain development. *Cell Stem Cell* **30**, 851-866.e7 (2023).
16. Bhaduri, A. *et al.* An atlas of cortical arealization identifies dynamic molecular signatures. *Nature* **598**, 200–204 (2021).
17. Hovestadt, V. *et al.* Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).
18. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017).
19. Northcott, P. A. *et al.* Medulloblastoma comprises four distinct molecular variants. *J. Clin. Oncol.* **29**, 1408–1414 (2011).
20. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
21. Northcott, P. A., Korshunov, A., Pfister, S. M. & Taylor, M. D. The clinical implications of medulloblastoma subgroups. *Nat. Rev. Neurol.* **8**, 340–351 (2012).
22. Ramaswamy, V. & Taylor, M. D. Medulloblastoma: From myth to molecular. *J. Clin. Oncol.* **35**, 2355–2363 (2017).
23. Gibson, P. *et al.* Subtypes of medulloblastoma have distinct developmental origins. *Nature* **468**, 1095–1099 (2010).
24. Hendrikse, L. D. *et al.* Failure of human rhombic lip differentiation underlies medulloblastoma formation. *Nature* **609**, 1021–1028 (2022).

25. Schüller, U. *et al.* Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma. *Cancer Cell* **14**, 123–134 (2008).

26. Yabe, T., Hata, T., He, J. & Maeda, N. Developmental and regional expression of heparan sulfate sulfotransferase genes in the mouse brain. *Glycobiology* **15**, 982–993 (2005).

27. Augustin, I. *et al.* The cerebellum-specific Munc13 isoform Munc13-3 regulates cerebellar synaptic transmission and motor learning in mice. *J. Neurosci.* **21**, 10–17 (2001).

28. Vanner, R. J. *et al.* Quiescent Sox2+ cells drive hierarchical growth and relapse in sonic hedgehog subgroup medulloblastoma. *Cancer Cell* **26**, 33–47 (2014).

29. Mullen, R. J., Buck, C. R. & Smith, A. M. NeuN, a neuronal specific nuclear protein in vertebrates. *Development* **116**, 201–211 (1992).

30. Gilthorpe, J. D., Papantoniou, E.-K., Chédotal, A., Lumsden, A. & Wingate, R. J. T. The migration of cerebellar rhombic lip derivatives. *Development* **129**, 4719–4728 (2002).

31. Sakakibara, S., Nakamura, Y., Satoh, H. & Okano, H. Rna-binding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian CNS. *J. Neurosci.* **21**, 8091–8107 (2001).

32. Chizhikov, V. V. *et al.* Lmx1a regulates fates and location of cells originating from the cerebellar rhombic lip and telencephalic cortical hem. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10725–10730 (2010).

33. Garancher, A. *et al.* NRL and CRX define photoreceptor identity and reveal subgroup-specific dependencies in medulloblastoma. *Cancer Cell* **33**, 435-449.e6 (2018).

34. Englund, C. *et al.* Unipolar brush cells of the cerebellum are produced in the rhombic lip and migrate through developing white matter. *J. Neurosci.* **26**, 9184–9195 (2006).

35. Forget, A. *et al.* Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. *Cancer Cell* **34**, 379-395.e7 (2018).

36. Northcott, P. A. *et al.* Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. *Nat. Genet.* **41**, 465–472 (2009).
37. Haldipur, P. *et al.* Evidence of disrupted rhombic lip development in the pathogenesis of Dandy-Walker malformation. *Acta Neuropathol.* **142**, 761–776 (2021).
38. Monje, M. *et al.* Hedgehog-responsive candidate cell of origin for diffuse intrinsic pontine glioma. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4453–4458 (2011).
39. Jacquet, B. V. *et al.* FoxJ1-dependent gene expression is required for differentiation of radial glia into ependymal cells and a subset of astrocytes in the postnatal brain. *Development* **136**, 4021–4031 (2009).
40. Betancourt, J., Katzman, S. & Chen, B. Nuclear factor one B regulates neural stem cell differentiation and axonal projection of corticofugal neurons. *J. Comp. Neurol.* **522**, 6–35 (2014).
41. Curry, R. N. *et al.* Glioma epileptiform activity and progression are driven by IGSF3-mediated potassium dysregulation. *Neuron* **111**, 682-695.e9 (2023).
42. Malatesta, P., Appolloni, I. & Calzolari, F. Radial glia and neural stem cells. *Cell Tissue Res.* **331**, 165–178 (2008).
43. Pollen, A. A. *et al.* Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015).
44. Sanai, N., Alvarez-Buylla, A. & Berger, M. S. Neural stem cells and the origin of gliomas. *N. Engl. J. Med.* **353**, 811–822 (2005).
45. Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumors. *Cancer Res.* **63**, 5821–5828 (2003).
46. Galli, R. *et al.* Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res.* **64**, 7011–7021 (2004).
47. Kessaris, N. *et al.* Competing waves of oligodendrocytes in the forebrain and postnatal elimination of an embryonic lineage. *Nat. Neurosci.* **9**, 173–179 (2006).

48. Wamsley, B. & Fishell, G. Genetic and activity-dependent mechanisms underlying interneuron diversity. *Nat. Rev. Neurosci.* **18**, 299–309 (2017).

49. Miyoshi, G., Butt, S. J. B., Takebayashi, H. & Fishell, G. Physiologically distinct temporal cohorts of cortical interneurons arise from telencephalic Olig2-expressing precursors. *J. Neurosci.* **27**, 7786–7798 (2007).

50. Benamer, N., Vidal, M. & Angulo, M. C. The cerebral cortex is a substrate of multiple interactions between GABAergic interneurons and oligodendrocyte lineage cells. *Neurosci. Lett.* **715**, 134615 (2020).

51. Fruttiger, M. *et al.* Defective oligodendrocyte development and severe hypomyelination in PDGF-A knockout mice. *Development* **126**, 457–467 (1999).

52. Curry, R. N. *et al.* Spiking GABAergic OPC tumor cells prolong survival in IDH1 mutant glioma. *bioRxiv* 2024.03.02.583026 (2024) doi:10.1101/2024.03.02.583026.

53. Lu, Q. R. *et al.* Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection. *Cell* **109**, 75–86 (2002).

54. Stolt, C. C. *et al.* Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. *Genes Dev.* **16**, 165–170 (2002).

55. Vallstedt, A., Klos, J. M. & Ericson, J. Multiple dorsoventral origins of oligodendrocyte generation in the spinal cord and hindbrain. *Neuron* **45**, 55–67 (2005).

56. Zhou, Q. & Anderson, D. J. The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. *Cell* **109**, 61–73 (2002).

57. Zhou, Q., Wang, S. & Anderson, D. J. Identification of a novel family of oligodendrocyte lineage-specific basic helix-loop-helix transcription factors. *Neuron* **25**, 331–343 (2000).

58. Bai, X., Kirchhoff, F. & Scheller, A. Oligodendroglial GABAergic signaling: More than inhibition! *Neurosci. Bull.* **37**, 1039–1050 (2021).

59. Hatakeyama, J. *et al.* Hes genes regulate size, shape and histogenesis of the nervous system by control of the timing of neural stem cell differentiation. *Development* **131**, 5539–5550 (2004).

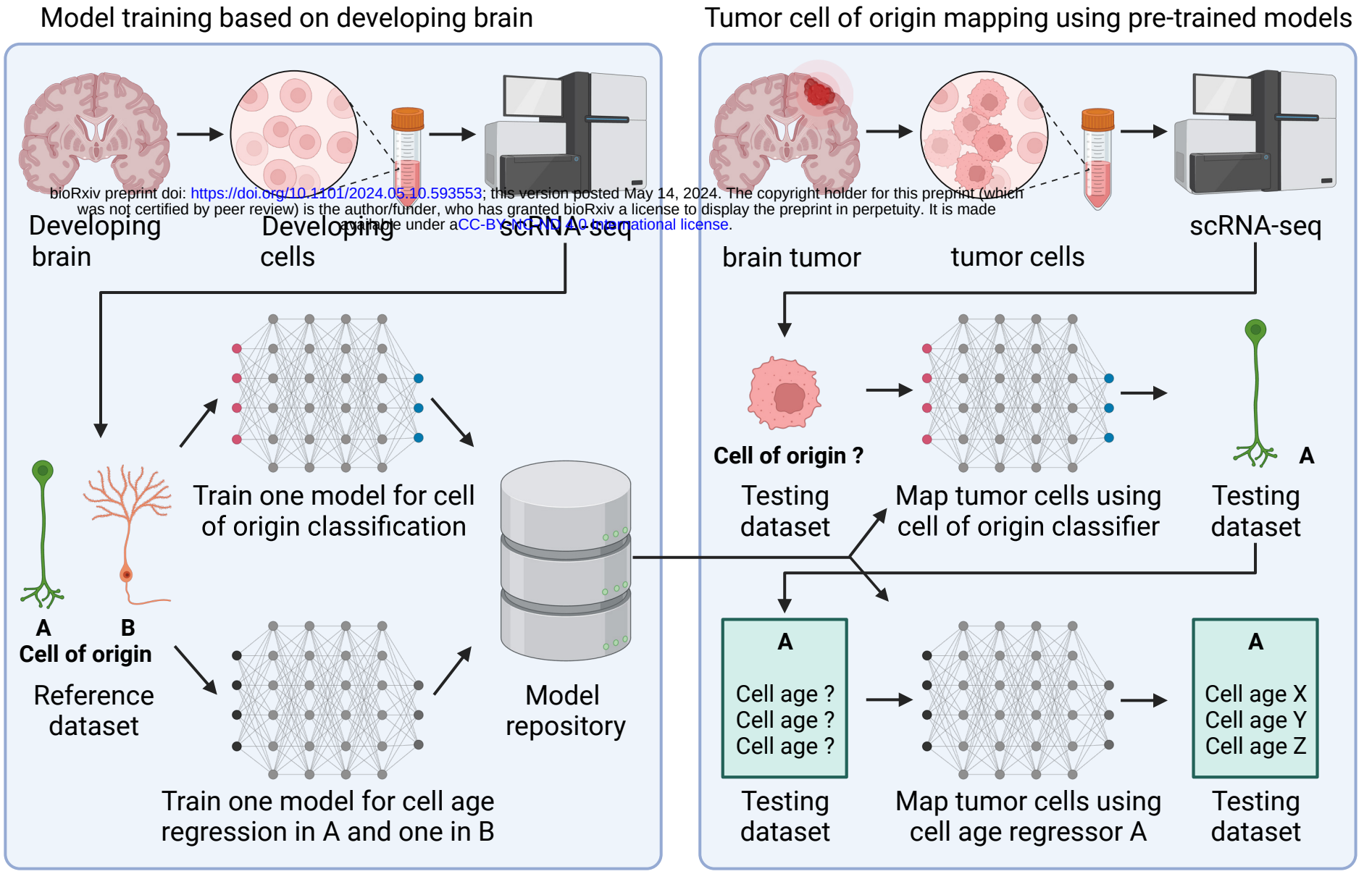
60. Graham, V., Khudyakov, J., Ellis, P. & Pevny, L. SOX2 functions to maintain neural progenitor identity. *Neuron* **39**, 749–765 (2003).

61. Kim, J. *et al.* Ttyh1 regulates embryonic neural stem cell properties by enhancing the Notch signaling pathway. *EMBO Rep.* **19**, e45472 (2018).

62. Yan, M. & Liu, Q. Differentiation therapy: a promising strategy for cancer treatment. *Chin. J. Cancer* **35**, (2016).

63. Hugues de Thé. Differentiation therapy revisited. *Nat. Rev. Cancer* **18**, 117–127 (2018).

A



B